Master's Thesis

Scribble2Label: Self-labeling via Consistency for Scribble-supervised Cell Segmentation

Hyeon-Soo Lee

Department of Computer Science and Engineering

Ulsan National Institute of Science and Technology

2021

Scribble2Label: Self-labeling via Consistency for Scribble-supervised Cell Segmentation

Hyeon-Soo Lee

Department of Computer Science and Engineering

Ulsan National Institute of Science and Technology

Scribble2Label: Self-labeling via Consistency for Scribble-supervised Cell Segmentation

A thesis submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Hyeon-Soo Lee

12/15/2020 of submission

Approved by

Advisor
Se-Young Chun

Scribble2Label: Self-labeling via Consistency for Scribble-supervised Cell Segmentation

Hyeon-Soo Lee

This certifies that the thesis of Hyeon-Soo Lee is approved.

12/15/2020 of submission

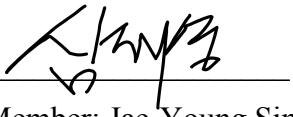
Signature

Advisor: Se-Young Chun

Signature

Committee Member: Won-ki Jeong

Signature



Committee Member: Jae-Young Sim

Signature

Abstract

Cell segmentation gives important findings in medical image analysis. Through cell analysis, various tasks such as cancer diagnosis, reconstruction of synaptic connectivity maps, measurement of drug response and so on could be possible.

With the advent of recent advances in deep learning, more accurate and high-throughput cell segmentation has become feasible. However, deep learning-based cell segmentation faces a problem of cost and scalability for constructing dataset. Supervised-learning methods require fully annotated ground-truth labels, where there are as many as hundreds of cells. Consequently, it needs time-consuming and labor-intensive works.

In this thesis, `Scribble2Label`, a novel weakly-supervised cell segmentation framework that exploits only a handful of scribble annotations without full segmentation labels. The core idea is to combine pseudo-labeling and label filtering to generate reliable labels from weak supervision. For this, we leverage the consistency of predictions by iteratively averaging the predictions to improve pseudo labels.

The performance of `Scribble2Label` is demonstrated by comparing it to several state-of-the-art cell segmentation methods with various cell image modalities, including bright-field, fluorescence, and electron microscopy. Our method achieves outperformed results compared with previous related works from various data including fluorescence, histopathology, Bright-field and electron microscopy(EM). Furthermore, the prop method consistently works well in different scribble instance levels.

Contents

List of Figures

I Introduction

1.1 Problem Definition

Micro- to nano-scale microscopy images are commonly used for cellular-level biological image analysis. In cell image analysis, segmentation serves as a crucial task to extract the morphology of the cellular structures. The applications of cell analysis have a wide range of the use, as Figure 1 shows. For example, in histopathology, cancer is diagnosed by cell morphological features [1]. In connectomics, neural circuits are reconstructed by segment cell images [2]. Drug response is also measurable by comparing the cell morphology before and after an injection [3].

Conventional cell segmentation methods are mostly grounded in model-based and energy minimization methods, such as Watershed [6], Chan-Vese with the edge model [7], and gradient vector flow [8]. The recent success of deep learning has gained much attention in many image processing and computer vision tasks. A common approach to achieve highly-accurate segmentation performance is to train deep neural networks using ground-truth labels [9–11]. However, generating a sufficient number of ground-truth labels is time-consuming and labor-intensive,

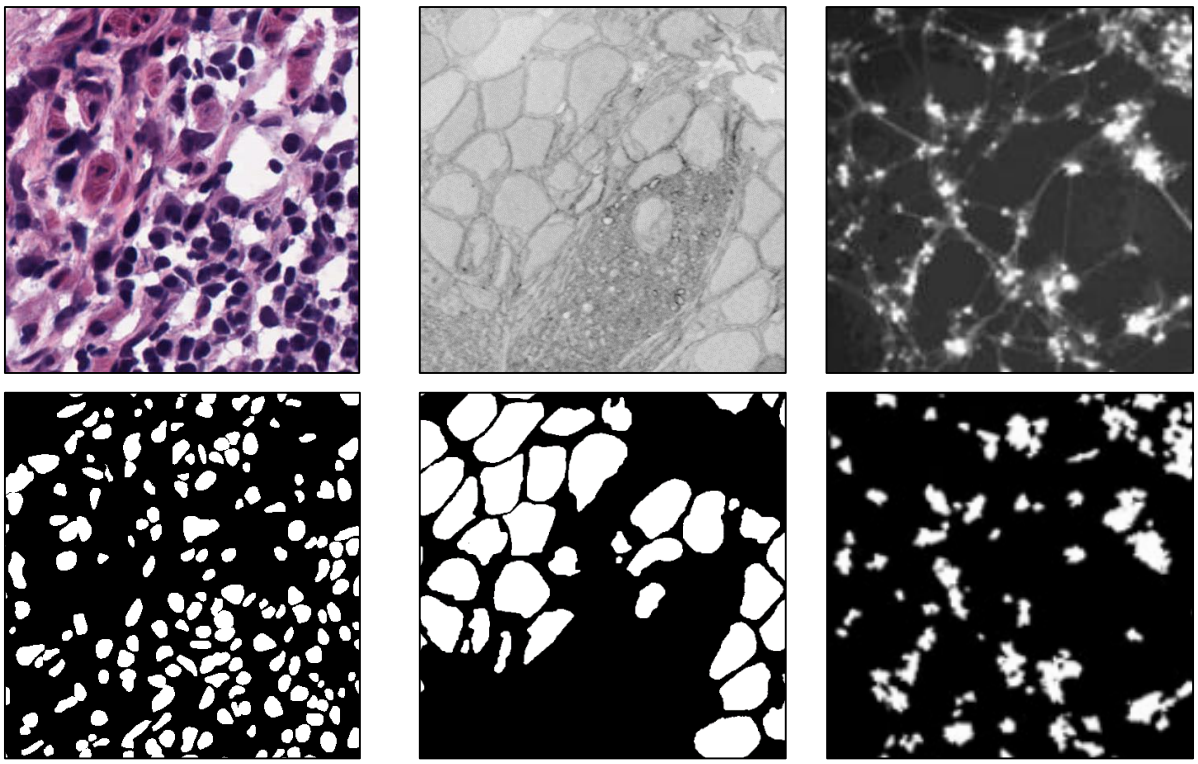


Figure 1: Applications of cell segmentation. The application examples of cell segmentation. (a) Cancer diagnosis by analyzing cells in a histopathology image [1]. (b) Reconstruction of synaptic connectivity maps by neuronal cell segmentation [2]. (c) Drug response measurement by cell morphological analysis [3].

which is becoming a major bottleneck in the segmentation process. Additionally, manually generated segmentation labels are prone to errors due to the difficulty in drawing pixel-level accurate region masks.

To address such problems, weakly-supervised cell segmentation methods using point annotation have recently been proposed [12–14]. Yoo et al. [14] and Qu et al. [13] introduced methods that generate coarse labels only from point annotations using a Voronoi diagram. Further, Nishimura et al. [12] proposed a point detection network in which output is used for cell instance segmentation. Even though point annotation is much easier to generate compared to full region masks, the existing work requires point annotations for the entire dataset – for example, there are around 22,000 nuclei in 30 images of the MoNuSeg dataset [4]. Moreover, the performance of the work mentioned above is highly sensitive to the point location, i.e., the point should be close to the center of the cell.

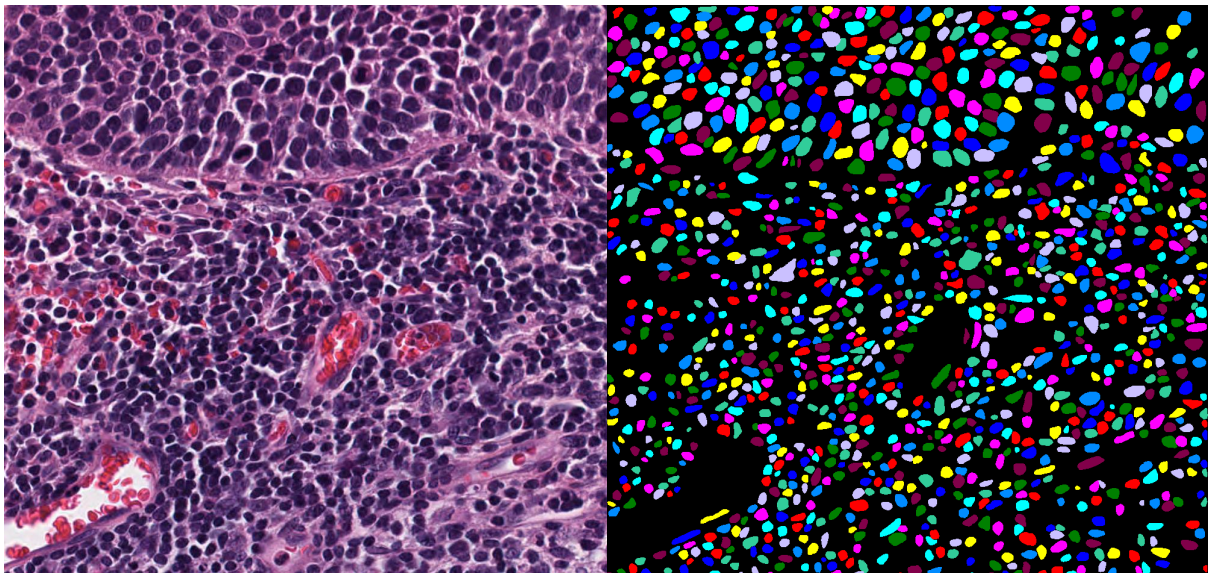
Recently, weakly-supervised learning using scribble annotations, i.e., scribble-supervised learning, has actively been studied in image segmentation as a promising direction for lessening the burden of manually generating training labels. Scribble-supervised learning exploits scribble labels and regularized networks with standard segmentation techniques (e.g., graphcut [15], Dense Conditional Random Field [DenseCRF] [16,17]) or additional model parameters (e.g., boundary prediction [18] and adversarial training [19]). The existing scribble-supervised methods have demonstrated the possibility to reduce manual efforts in generating training labels, but their adaptation in cell segmentation has not been explored yet.

1.2 Motivation

With the advent of recent advances in deep learning, image segmentation has made huge improvement. Although the existing image segmentation method used energy-based segmentation methods, it was difficult to operate robustly on images taken in various environments. However, through a convolutional natural network, it was able to represent meaningful features from an image and the generalization performance is greatly improved in the image recognition task [20].

Recently, however, the high cost of building datasets has been one of the image segmentation bottlenecks. A lot of research have been done actively to address this problem. Various methods have been proposed to reduce the cost of datasets by using a relatively inexpensive label, such as image-level labels [21, 22], scribble labels [15, 16, 16–19], point labels [13, 14], and partial labels [23–25], with minimal segmentation performance decline. Cell segmentation is especially expensive for building datasets. In Fig. 2, there are hundreds of cells in a single medical image. Generating hundreds of cell label while keeping their boundaries complete shape takes a lot of concentration and time. In addition, due to the characteristics of medical images, only experts can annotate, which is expensive.

To address this problem, we have studied how the neural network can effectively recognize the semantic information of the image using only scribble, one of the intuitive label types that has been used for a long time. In the existing natural image, the segmentation was carried out by utilizing scribble through graphed-based algorithms [15–17, 19]. However, we focused to solve the high cost problem of deep learning by taking full advantage of the training process of the neural network without additional computing.



(a) (b)

Figure 2: A sample image and label from [4]. There are 1,309 cells in this single 1,000 x 1,000 pathology image.

1.3 Contribution

In this thesis, we propose a novel weakly-supervised cell segmentation method that is highly accurate and robust with only a handful of manual annotations. Our method, `Scribble2Label`, uses scribble annotations as in conventional scribble-supervised learning methods, but we propose the combination of pseudo-labeling and label-filtering to progressively generate full training labels from a few scribble annotations. By doing this, we can effectively remove noise in pseudo labels and improve prediction accuracy. The main contributions of our work are as follows.

- We introduce a novel iterative segmentation network training process that generates training labels automatically via weak-supervision using only a small set of manual scribbles, which significantly reduces the manual effort in generating training labels.
- We propose a novel idea of combining pseudo-labeling with label filtering, exploiting consistency to generate reliable training labels, which results in a highly accurate and robust performance. We demonstrate that our method consistently outperforms the existing state-of-the-art methods across various cell image modalities and different levels of scribble details.
- Unlike existing scribble-supervised segmentation methods, our method is an end-to-end scheme that does not require any additional model parameters or external segmentation methods (e.g, Graph-cut, DenseCRF) during training.

To the best of our knowledge, this is the first scribble-supervised segmentation method applied to the cell segmentation problem in various microscopy images.

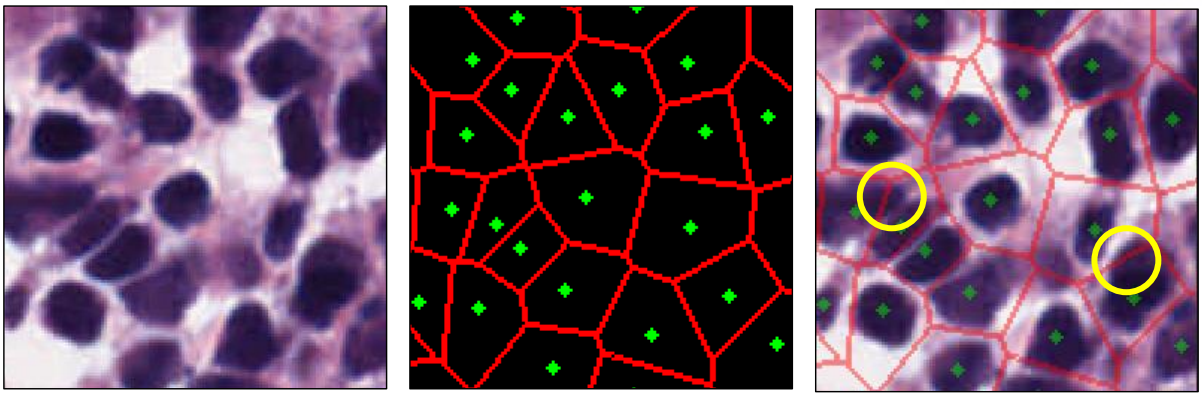
II Background

2.1 Cell Segmentation

Cell segmentation is an important step in cell imaging analysis and has been steadily being studied in the field of computer vision. Methods vary: Segmentation based on Watershed [6], gradient vector flow [8] and Chan-vese with edge model [7].

Watershed is one of the popular approach to segment cells. It is the energy-based segmentation method, which separates the image parts by assuming an image brightness as geological height [26]. However, edge recognition is a challenge with Watershed because only the brightness is the consideration. Lots of noise are existed in medical images depending on time and environment. Local observations on textures can't help to divide a cell instance because noise disturb the boundary of cells. To address this problem, edge-preserving segmentation methods are proposed such as gradient vector flow and Chan-vese with edge model. [8] designed edge-preserve gradient vector flow so that the snakes can be stopped even along a weak boundaries. [7] proposed Chan-vese based cell segmentation method on fluorescence time-lapse series images. This method accomplished instance-aware function by utilizing the characteristics of time-lapse and fluorescence images which neighbour frames are closely related. But, these energy-based method are highly affected by manual parameters. This property limits segmentation techniques to the various environments.

Since the introduction of deep learning, the generalization of cell segmentation has improved significantly. [9] exploits a convolutional neural network(CNN) to detect cells and selection-based shape model to separate overlapped cells. With high-level feature from CNN, this approach can be applied on various staining conditions. [10] proposed U-Net [27] combined with Long Short-Term Memory(LSTM) [28] to aggregate time sequence information in live cell microscopy sequences. This method also exploits the boundary-aware loss to segment cells which are tracing through time sequences via LSTM-combined convolution modules. Object detection network is also utilized to perform the instance cell segmentation task. [11] proposed a instance segmentation network by generating proper bounding boxes with keypoint graph grouping. Previous object detection works than [11] were hard to segment cells because of the limitation from anchor boxes proposal. It is challenging to propose dense bounding boxes where cells are cluttered, because non maximum suppression(NMS) operation aggregates the small overlapped boxes. But the keypoint graph based detection network is freed from this challenge. However, most of the deep learning based cell segmentation methods [9–11] require a fully annotated label. Especially, a medical image can be labeled by experts and a cell image contains a lots of objects. Labeling sophisticated cell boundaries from cells requires the high-concentration. This is time-consuming to make and can be done by professionals, which is expensive.



(a) Image (b) Voronoi Edge (c) Overlapped Image

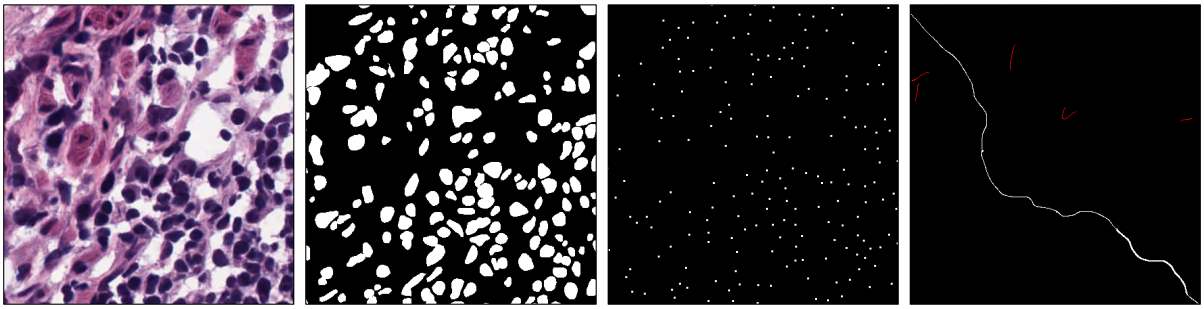
Figure 3: Example of Voronoi Edge. In (c), some parts of Voronoi Edge present upon the elongated cells.

2.2 Weakly-supervised Cell Segmentation

Recently, weakly-supervised cell segmentation has been studied. Annotating all areas of a cell requires a lot of time and expert knowledge. To solve this problem, cell segmentation was performed using a point annotation.

[13, 14] used Voronoi edge to get the texture information of a cell and background. Voronoi edge also gives a cue to generate a new coarse label. In [13], a new coarse label are created from Voronoi edge by clustering and used to segment cells in histopathology images. [14] not only exploits Voronoi edge to make a network train about a cell and background, but also let an auxiliary network recognize a boundary by giving segmentation result's edges. In [12], the point detection network is trained using point annotation, and the cell instance segmentation is performed by using the image area that affects the backpropagation of the center point of the detected cell.

However, there are still too many nuclei in a cell image. All of cells are required to be marked and this work asks an annotator concentrating hard on labeling. Moreover, one more thing to be concentrated is the point's location. The location of points affects the performance of point-based methods. Annotating the center of many cells requires mental work. In addition, as in Fig. 3, the part of Voronoi edge presents upon an elongated cell where it is considered as a background label. This noisy labels can cause a network's performance degradation. So a scribble-supervised cell segmentation algorithm, that learns by simply annotating a small number of cells and backgrounds with a scribble, is proposed.



(a) Image (b) Full Label (c) Point Label (d) Scribble Label

Figure 4: Example of different label types

2.3 Scribble-supervised Learning

Scribble is one of the simple annotation methods. Scribble is widely used in interactive image segmentation [15]. Scribble-supervised Learning is a way to train a model with only scribble labels. Scribble labels requires a relatively small amount of labor compared to the point labels. Because the point is to mark every cell one by one, whereas the scribble is intuitively annotated with simple lines for background and cells.

Scribble-supervised Learning exploits scribble-label itself and regularized the network with standard segmentation technique or additional parameters [15–19]. Standard energy-based segmentation methods are combined with the deep-learning training process. [15] regularizes the proposal generation by graph-cut. [16] proposed the label initialization from a scribble label and re-labeling process by using Random Walk and DenseCRF. [17] integrates DenseCRF into the training loss directly to give correlation information. However, using standard segmentation processes such as Random Walk and DenseCRF increases the computation costs.

In another direction, the additional model parameters are also helpful for improving scribble supervised learning, such as boundary prediction [18] and discriminator [19]. [18] creates another branch to predict the boundary of objects in a segmentation network. It improves segmentation performance by making the network directly predict edge, which is insufficient information in the scribble label. [19] proposed an adversarial learning to train a network with a handful of scribbles. This method generates a bounding box label from scribbles by calculating principal component analysis(PCA) and let a segmentation network try to predict the bounding box. Scribble label also gives semantic information to the segmentation network.

However, previous works exploits additional energy-based segmentation process or additional network. These supplementary computations increase time cost and expenses. For example, DenseCRF is one of widely used methods for giving semantic information. But, in conventional formulations of CRFs, the computation cost is exponentially increased by the number of previous states [29]. If all pixels are considered, where the pixels would be the states, the computational cost is extremely expensive. Another directions, which exploit multi-task learning or adversarial learning, also rise expenses in terms of time and memory.

Unlike previous methods, `Scribble2Label` doesn't require additional standard segmentation computation or model parameters. Proposed method is motivated not only to reduce time and computational cost, but also to use well semantic information from scribbles. The network generates a reliable label from scribble by combining pseudo-labeling and label-filtering. The details of proposed method is covered in Section III.

2.4 Semi-supervised Learning

Semi-supervised Learning (SSL) is a method of training a model using a few labeled data and a large number of unlabeled data. The general loss in SSL is defined as,

$$\mathcal{L} = \mathcal{L}_{\chi} + \lambda_{\mathcal{U}} \mathcal{L}_{\mathcal{U}}, \tag{1}$$

where $\mathcal{L}$ is a general loss function combined with supervised loss $\mathcal{L}_{\chi}$ and semi-supervised loss $\mathcal{L}_{\mathcal{U}}$, and $\lambda_{\mathcal{U}}$ is a loss weight for $\mathcal{L}_{\mathcal{U}}$. The purpose of $\mathcal{L}_{\mathcal{U}}$ is the regularization of a network with unlabeled data.

There are several ways to define $\mathcal{L}_{\mathcal{U}}$, such as consistency regularization with data augmentation, entropy minimization and traditional regularization [30]. Data augmentation is one of the regularization techniques to prevent overfitting in a neural network. In supervised learning, there is the assumption where class distribution of prediction should not be affected by data augmentation. But, in semi-supervised learning, this concept applies that class distribution must be constant even if the data is augmented. This technique is called as consistency regularization. Recent work [31] exploits the consistency regularization by force a network to output the similar class distribution from strongly-augmented data with weakly-augmented data.

In SSL, entropy minimization generally enforces that the classifier's decision boundary doesn't pass the high-density regions of the marginal data distribution [30]. Entropy minimization makes output low-entropy prediction from the unlabeled data. The empirical distribution of unlabeled can be assumed as the confidence of prediction should be high. To maximize this empirical distribution, that is, the maximizer of the the posterior distrubtion,

$$\mathcal{L}_{\mathcal{U}} = \lambda_{\mathcal{U}} \sum_{i}^{n_{\mathcal{U}}} f(x_i;\theta) log f(x_i;\theta) \tag{2}$$

where x is an input image, $n_{\mathcal{U}}$ is the number of unlabeled data. $f(x;\theta)$ is the model's prediction. [31, 32] utilize pseudo-labeling to make one-hot encoding a sample with high prediction confidence. Psuedo-Label is defined as,

$$\hat{y} = \mathbb{1}(max(f(x;\theta)) > \tau), \tag{3}$$

where τ is the confident threshold. [30] makes the sharpen target distribution, which is closer to the one-hot distribution. [25] uses pseudo-labeling to generate confidence values from discriminators.

Regularization has a benefit of preventing the memorization of a deep neural network. To achieve this functionality, the previous work [33], proposed the data augmentation technique which mixes two different data and their labels called as MixUp. MixUp is used to generate a label for unlabeled data in SSL [30]. Furthermore, semi-supervised segmentation [23] allows a network to train through strong permutation by cropping and mixing different pairs of data which is motivated by CutMix [34] in image classification task.

Conventional Pseudo-label can be a biased distribution based on a specific state in a network that is different from the actual ground truth. In this thesis, `Scribble2Label` suggests a pseudo-labeling method that collaborates with the consistency of prediction.

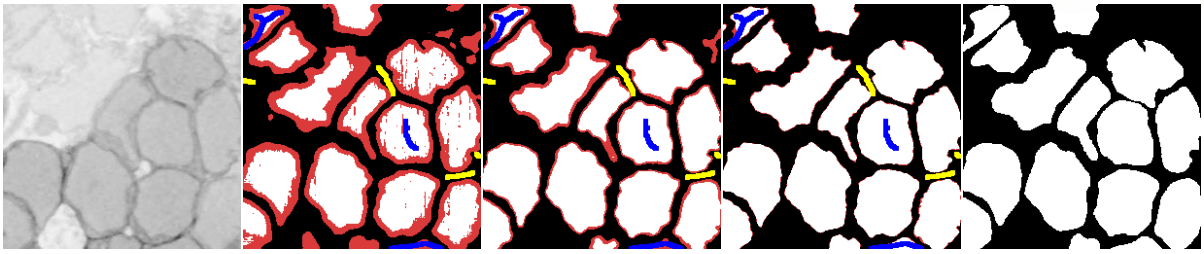


Figure 5: An example of iterative refinement of pseudo labels during training. Blue and yellow: scribbles for cells and background, respectively (Ω_s); red: the pixels below the consistency threshold τ, which will be ignored when calculating the unscribbled pixel loss ($\mathcal{L}_{up}$); white and black: cell or background pixels over τ (Ω_g). (a) – (c) represent the filtered pseudo-labels from the predictions over the iterations (with Intersection over Union [IoU] score): (a): 7th (0.5992), (b): 20th (0.8306), and (c): 100th (0.9230). The actual scribble thickness used in our experiment was 1 pixel, but it is widened to 5 pixels in this figure for better visualization.

III Method

In this section, the proposed segmentation method is described in detail. The input sources for our method are the image x and the user-given scribbles s (see Figure 6). Here, the given scribbles are *labeled* pixels (denoted as blue and yellow for the foreground and background, respectively), and the rest of the pixels are *unlabeled* pixels (denoted as black). For labeled (scribbled) pixels, a standard cross-entropy loss is applied. For unlabeled (unscribbled) pixels, our network automatically generates reliable labels using the exponential moving average of the predictions during training. Training our model consists of two stages. The first stage is initialization (i.e., a warm-up stage) by training the model using only the scribbled pixel loss ($\mathcal{L}_{sp}$). Once the model is initially trained via the warm-up stage, the prediction is iteratively refined by both scribbled and unscribbled losses ($\mathcal{L}_{sp}$ and $\mathcal{L}_{up}$). Figure 6 illustrates the overview of the proposed system.

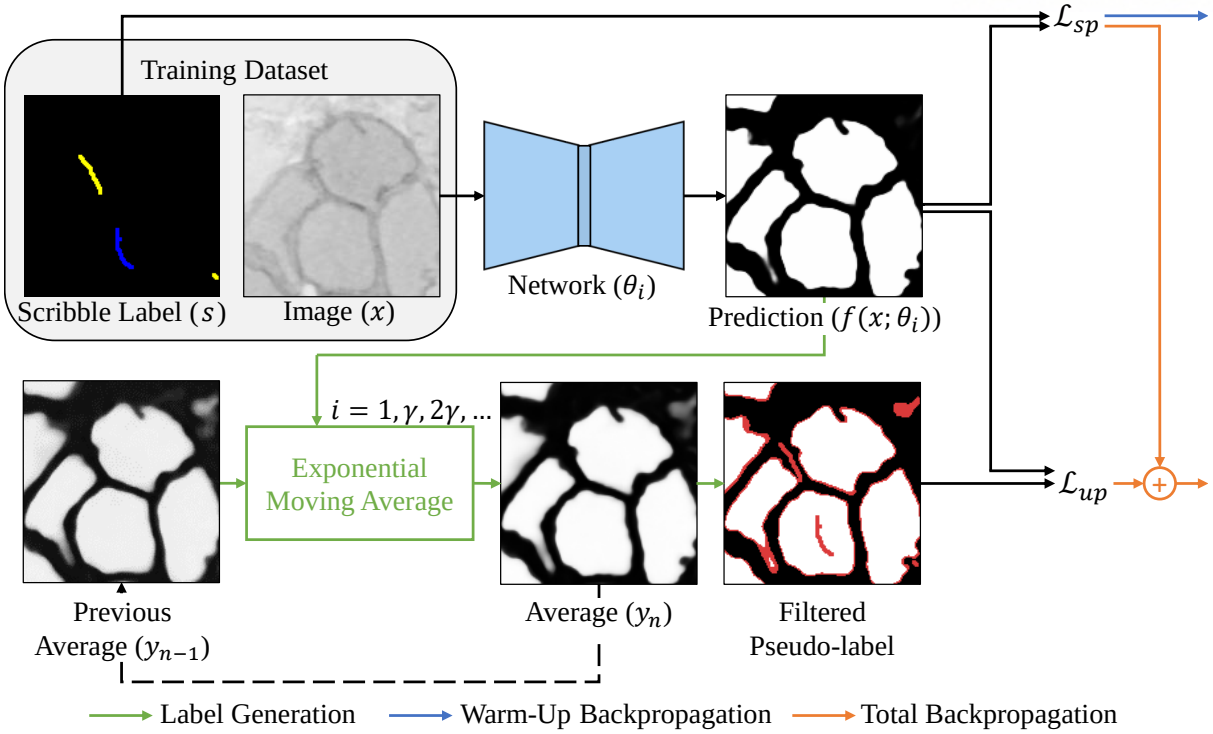


Figure 6: The overview of the proposed method (`Scribble2Label`). The pseudo-label is generated from the average of predictions. Following, $\mathcal{L}_{sp}$ is calculated with the scribble annotation, and $\mathcal{L}_{up}$ is calculated with the filtered pseudo-label. The prediction ensemble process occurs every γ epochs, where γ is the ensemble interval. n represents how many times the predictions are averaged.

3.1 Warm-Up Stage

At the beginning, we only have a small set of user-drawn scribbles for input training data. During the first few iterations (warm-up stage), we train the model only using the given scribbles, and generate the average of predictions which can be used in the following stage (Section 3.2). Here, the given scribbles is a subset of the corresponding mask annotation. By ignoring unscribbled pixels, the proposed network is trained with cross entropy loss as follows:

$$\mathcal{L}_{sp}(x, s) = -\frac{1}{|\Omega_s|} \sum_{j \in \Omega_s} [s_j \log(f(x;\theta_i)) + (1 - s_j) \log(1 - f(x;\theta_i))], \tag{4}$$

where x is an input image, s is a scribble annotation, and Ω_s is a set of scribbled pixels. $f(x;\theta_i)$ is the model's prediction at iteration i. This warm-up stage continues until we reach the warm-up Epoch E_W.

Moreover, we periodically calculate the exponential moving average (EMA) of the predictions over the training process: $y_n = \alpha f(x;\theta_i) + (1 - \alpha) y_{n-1}$ where α is the EMA weight, y is the average of predictions, $y_0 = f(x;\theta_1)$, and n is how many times the predictions are averaged. This process is called a prediction ensemble [35]. Note that, since we use data augmentation for training, the segmentation prediction is not consistent for the same input image. Our solution for

this problem is splitting the training process into training and ensemble steps. In the ensemble phase, an un-augmented image is used for the input to the network, and EMA is applied to that predictions. Moreover, in the scribble-supervised setting, we cannot ensemble the predictions when the best model is found, as in [35], because the given label is not fully annotated. To achieve the valuable ensemble and reduce computational costs, the predictions are averaged every γ epochs, where γ is the ensemble interval.

3.2 Learning with a Self-Generated Pseudo-Label

The average of the predictions can be obtained after the warm-up stage. This can be used as a label for unscirbbled pixels. However, this average itself is noisy because it comes from the model's prediction. Thus, we need to filter it. For filtering the pseudo-label, the average is used. The pixels with consistently the same result are one-hot encoded and used as a label for unscribbled pixels with standard cross entropy. Using only reliable pixels and making these one-hot encoded progressively provide benefits through curriculum learning and entropy minimization [31]. With filtered pseudo-label, the unscribbled pixel loss is defined as follows:

$$\mathcal{L}_{up}(x, y_n) = -\frac{1}{|\Omega_g|} \sum_{j \in \Omega_g} [\mathbb{1}(y_n > \tau) \log(f(x; \theta_i)) + \mathbb{1}((1 - y_n) > \tau)) \log(1 - f(x; \theta_i))], \quad (5)$$

where $\Omega_g = \{g | g \in (max(y_n, 1 - y_n) > \tau), g \notin \Omega_s\}$, which is a set of generated label pixels, and τ is the consistency threshold. Formally, at iteration i, $\mathcal{L}_{up}$ is calculated with (x, y_n), where $n = \lfloor i/\gamma \rfloor + 1$. For unscribbled pixels, cross entropy is calculated by a prediction and pseudo-label from the prediction ensemble. The total loss is then defined as the combination of the scribbled loss $\mathcal{L}_{sp}$ and the unscribbled loss $\mathcal{L}_{up}$ with the relative weight of $\mathcal{L}_{up}$, defined as follows:

$$\mathcal{L}_{total}(x, s, y_n) = \mathcal{L}_{sp}(x, s) + \lambda \mathcal{L}_{up}(x, y_n) \quad (6)$$

In total, we combined $\mathcal{L}_{sp}$ and $\mathcal{L}_{up}$ to train a network. λ sets the relative weight of $\mathcal{L}_{up}$. Note the EMA method shown above is also applied during this training process. Algorithm 1 describes full processes.

Algorithm 1: Pseudo-code of `Scribble2Label`

Data: Training data $(x, s) \in \mathcal{D}$; Model Parameter θ; Warm-up Epoch E_W; Total Epoch E_T; Consistency Threshold τ; EMA Alpha α; Ensemble Interval γ; Average Count n;

Result: Model Parameter θ

$y_1 \leftarrow \emptyset$

$n \leftarrow 0$

`/* Warm-up Stage */`

for $i = 0, 1, 2, \ldots, E_W$ **do**

  $(\hat{x}, \hat{s}) = \text{Augment}(x, s)$

  Update θ_i by $\mathcal{L}_{sp}(\hat{x}, \hat{s})$

  if $MOD(i, \gamma) = 0$ **then**

    $n \leftarrow n + 1$

    $y_n \leftarrow \alpha f(x; \theta_i) + (1 - \alpha) y_{n-1}$

`/* Learning with Self-generated Pseudo-Label */`

for $i = E_W, \ldots, E_T$ **do**

  $(\hat{x}, \hat{s}, \hat{y_n}) = \text{Augment}(x, s, y_n)$

  Update θ_i by $\mathcal{L}_{total}(\hat{x}, \hat{s}, \hat{y_n})$

  if $MOD(i, \gamma) = 0$ **then**

    $n \leftarrow n + 1$

    $y_n \leftarrow \alpha f(x; \theta_i) + (1 - \alpha) y_{n-1}$

IV Results

4.1 Datasets

The efficacy of our method is demonstrated using three different cell image datasets. The first set, MoNuSeg [4], consists of 30 1000×1000 histopathology images acquired from multiple sites covering diverse nuclear appearances. A 10-fold cross-validation is conducted for the MoNuSeg dataset. BBBC038v1 [5], the second data set, which is known as Data Science Bowl 2018, is a set of nuclei 2D images. The stage 1 training dataset, which is fully annotated, is used and further divided into three main types, including 542 fluorescence (DSB-Fluo) images of various sizes, 108 320×256 histopathology images (DSB-Histo), and 16 bright-field 1000×1000 (DSB-BF) images.

Each dataset is split into training, validation, and test sets, with ratios of 60%, 20%, and 20%, respectively. EM is an internally collected serial-section electron microscopy image dataset of a larval zebrafish. There are three sub-volumes of either 512×512×512 or 512×512×256 in size. The size of the testing volume was 512×512×512.

The scribbles of MoNuSeg and DSBs were manually drawn by referencing the full segmentation labels. To ensure that the scribbles are generated without much efforts, the total time spent is restricted for each scribble annotation to one minute for images up to 256×256, two minutes for images up to 512×512, and four minutes for images up to 1024×1024 in size. For the EM dataset, the scribble annotation was generated by a scribble generation algorithm in [19] with a 10% ratio.

4.2 Implementation Details

Our baseline network was U-Net [27] with the ResNet-50 [20] encoder. For comparison with [13] in histopathology experiments (MoNuSeg, DSB-Histo), ResNet-34 is used for the encoder. The network was initialized with pre-trained parameters, and RAdam [36] was used for all experiments. In addition, we utilized the cosine annealing learning rate scheduler to give variability to the model parameters during the training process. To regularize the network, conventional data augmentation methods, such as cropping, flipping, rotation, shifting, scaling, brightness change, and contrast changes, are used.

The hyper-parameters used for our model are as follows: Consistency Threshold $\tau = 0.8$; EMA Alpha $\alpha = 0.2$; Ensemble Momentum $\gamma = 5$; $\mathcal{L}_{up}$'s weight $\lambda = 0.5$; and warm-up epoch $E_W = 100$. For the MoNuSeg dataset (which is much noisier than other datasets), $\tau = 0.95$ and $\alpha = 0.1$ to cope with noisy labels.

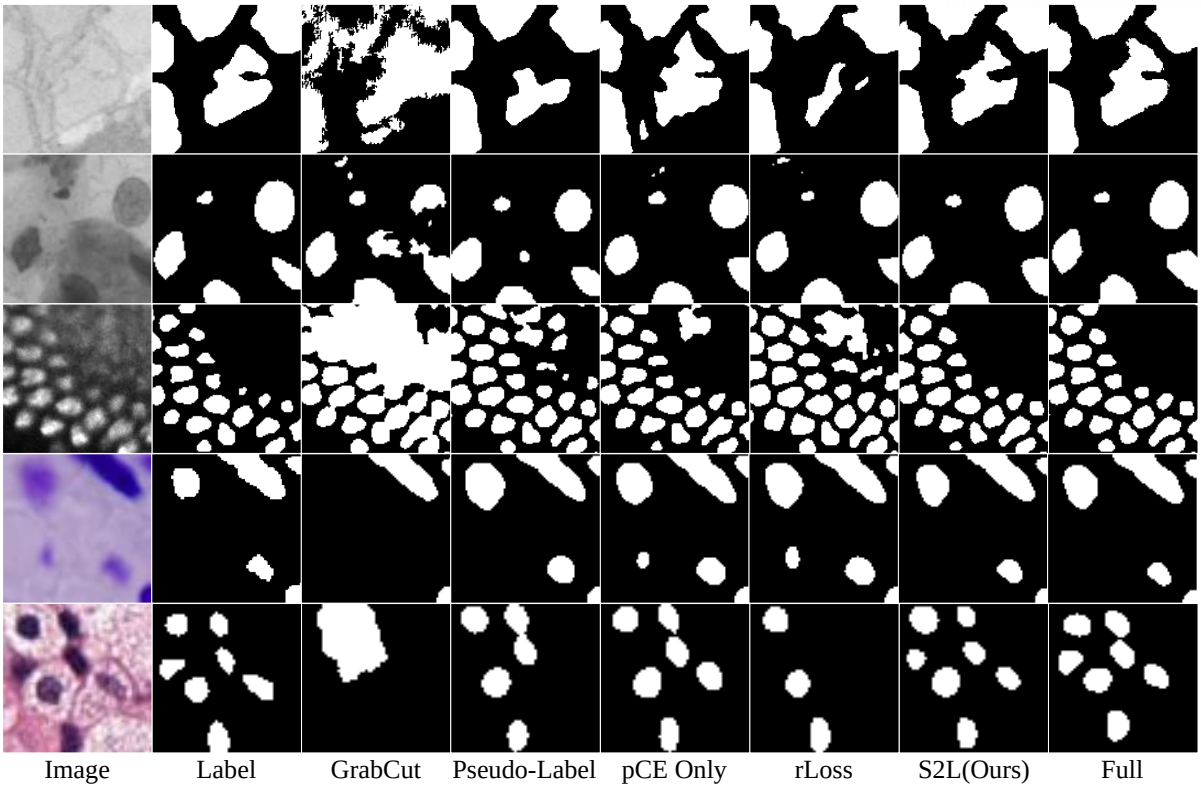


Figure 7: Qualitative results comparison. From the top to the bottom, EM, DSB-BF [5], DSB-Fluo, DSB-Histo, and MoNuSeg [4] are shown.

4.3 Results

The performance of semantic segmentation is evaluated using the intersection over union (IoU) and the performance of instance segmentation using mean Dice-coefficient (mDice) used in [12].

Comparison with other methods

The proposed method is compared to the network trained with full segmentation annotation, scribble annotation (pCE Only) [17], and the segmentation proposal from Grab-Cut [15]. To demonstrate the efficacy of the label filtering with consistency, it is compared to pseudo-labeling [32]. The pixels for which the probability of prediction were over threshold τ were assigned to be a pseudo-label, where τ was same as our method setting. Our method was also compared to Regularized Loss (rLoss) [17], which integrates the DenseCRF into the loss function. The hyper-parameters of rLoss are $\sigma_{XY} = 100$ and $\sigma_{RGB} = 15$.

Table 1 shows the quantitative comparison of our method with several representative methods. Overall, our method outperformed all methods on both IoU and mDice quality metrics. The proposed method achieved even higher mDice accuracy compared to the full method (i.e., trained using full segmentation labels) on EM, DSB-BF, and DSB-Histo datasets. Note also that MoNuSeg dataset contains many small cluttering cells, which are challenge to separate individually. However, our method showed outstanding instance segmentation results in this

Table 1: Quantitative results of various cell image modalities. The numbers represent accuracy in the format of IoU[mDice].

Label	**Method**	**EM**	**DSB-BF**	**DSB-Fluo**	**DSB-Histo**	**MoNuSeg**
Scribble	GrabCut [15]	0.5288 [0.6066]	0.7328 [0.7207]	0.8019 [0.7815]	0.6969 [0.5961]	0.1534 [0.0703]
	Pseudo-Label [32]	0.9126 [0.9096]	0.6177 [0.6826]	0.8109 [0.8136]	0.7888 [0.7096]	0.6113 [0.5607]
	pCE Only [17]	0.9000 [0.9032]	0.7954 [0.7351]	0.8293 [0.8375]	0.7804 [0.7173]	0.6319 [0.5766]
	rLoss [17]	0.9108 [0.9100]	0.7993 [0.7280]	0.8334 [0.8394]	0.7873 [0.7177]	0.6337 [0.5789]
	S2L(Ours)	**0.9208 [0.9167]**	**0.8236 [0.7663]**	**0.8426 [0.8443]**	**0.7970 [0.7246]**	**0.6408 [0.5811]**
Point	*Qu* [13]	-	-	-	0.5544 [0.7204]	0.6099 [0.7127]
Full	Full	0.9298 [0.9149]	0.8774 [0.7879]	0.8688 [0.8390]	0.8134 [0.7014]	0.7014 [0.6677]

case, too.

Grab-Cut's [15] segmentation proposal and the pseudo-label [32] were erroneous. Thus, training with these erroneous segmentation labels impairs the performance of the method. Qu et al.'s method [13] performed well for instance-level segmentation on MoNuSeg dataset, however, it performed worse on DSB-histo dataset. Because [13] used a clustering label that has circular shape cell label, it was hard to segment the non-circular cell. Learning with pCE [17] showed stable results on various datasets. However, due to learning using only scribbles, the method failed to correctly predict boundary accurately as in our method. rLoss [17] outperformed most of the previous methods, but our method generally showed better results. In terms of speed, our method is 30% faster than rloss training iteration, because it doesn't need to calculate the additional calculation process. We also observed that leveraging consistency by averaging predictions is crucial to generate robust pseudo-labels. Scribble2Label's results also confirm that using pseudo label together with scribbles is effective to generate accurate boundaries, comparable to the ground-truth segmentation label.

Table 2: Quantitative results using various amounts of scribbles. DSB-Fluo [5] was used for the evaluation. The numbers represent accuracy in the format of IoU[mDice].

Method	**10%**	**30%**	**50%**	**100%**	**Manual**
GrabCut [15]	0.7131 [0.7274]	0.8153 [0.7917]	0.8244 [0.8005]	0.8331 [0.8163]	0.8019 [0.7815]
Pseudo-Label [32]	0.7920 [0.8086]	0.7984 [0.8236]	0.8316 [0.8392]	0.8283 [0.8251]	0.8109 [0.8136]
pCE Only [17]	0.7996 [0.8136]	0.8180 [0.8251]	0.8189 [0.8204]	0.8098 [0.8263]	0.8293 [0.8375]
rLoss [17]	0.8159 [0.8181]	0.8251 [0.8216]	0.8327 [0.8260]	0.8318 [0.8369]	0.8334 [0.8394]
S2L(Ours)	**0.8274 [0.8188]**	**0.8539 [0.8407]**	**0.8497 [0.8406]**	**0.8588 [0.8443]**	**0.8426 [0.8443]**
Full	0.8688 [0.8390]				

Effect of amount of scribble annotations

To demonstrate the robustness of our method over various levels of scribble details, we conducted an experiment using scribbles automatically generated using a similar method by Wu et al. [19] (i.e., foreground and background regions are skeletonized and sampled). The target dataset was DSB-Fluo, and various amounts of scribbles, i.e., 10%, 30%, 50%, and 100% of the skeleton pixels extracted from the full segmentation labels (masks), are automatically generated. Table 2 summarizes the results with different levels of scribble details. Our method `Scribble2Label` generated stable results in both the semantic metric and instance metric from sparse scribbles to abundant scribbles.

The segmentation proposal from Grab-Cut [15] and the pseudo-lable [32] were noisy in settings lacking annotations, which resulted in degrading the performance. rLoss [17] performed better than the other methods, but it sometimes failed to generate correct segmentation results especially when the background is complex (causing confusion with cells). Our method showed very robust results over various scribble amounts. Note that our method performs comparable to using full segmentation masks only with 30% of skeleton pixels.

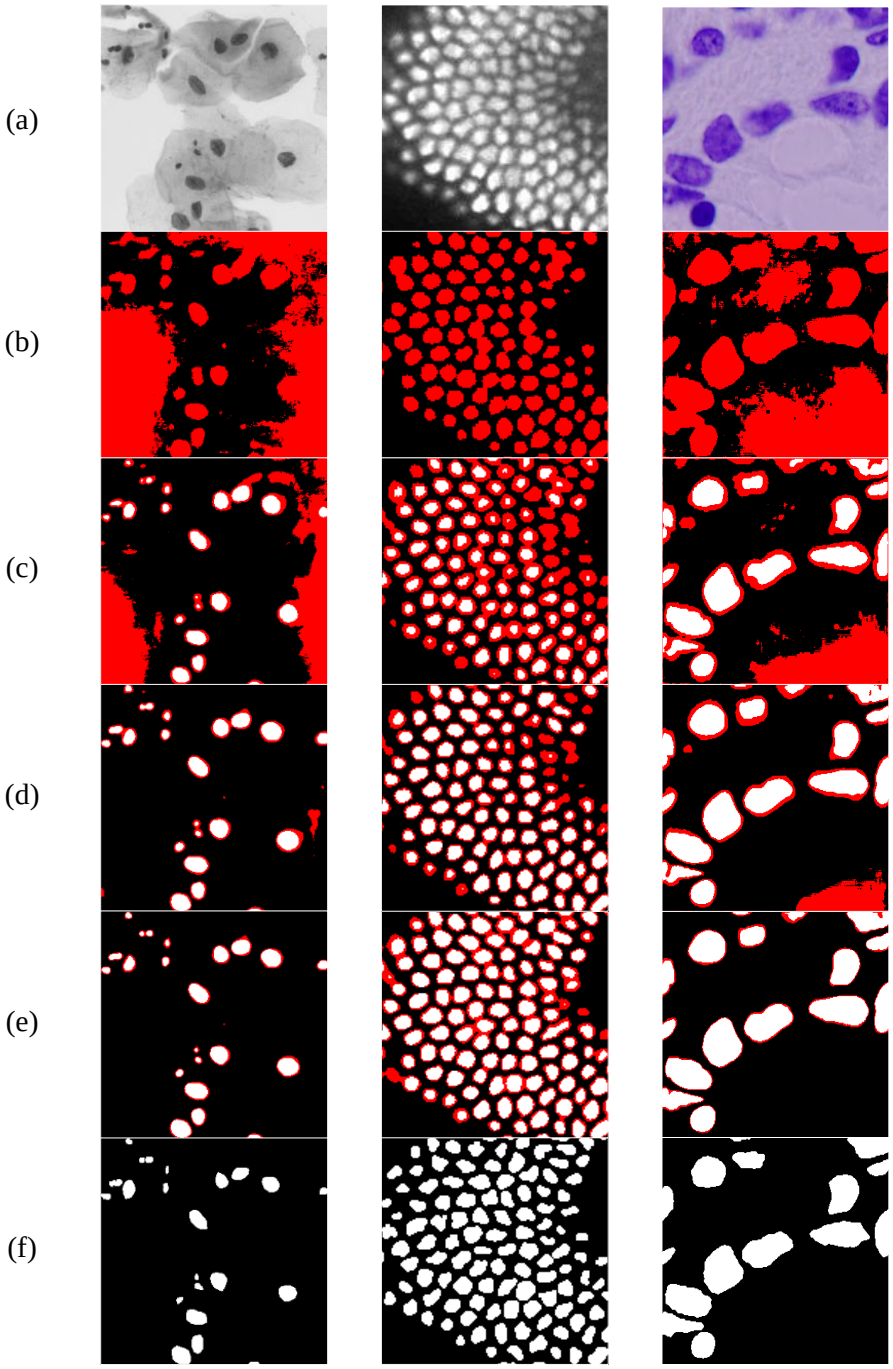


Figure 8: The self-label generation process from a small set of scribbles. The red is the pixel below the consistency threshold τ, the white and black are the cell or background pixel over τ. (a) is an input image, (b)-(e) are the label generation results as training progresses, (f) is a full label. We can observe the self-generated label gets close to ground truth label as pseudo-label is purified.

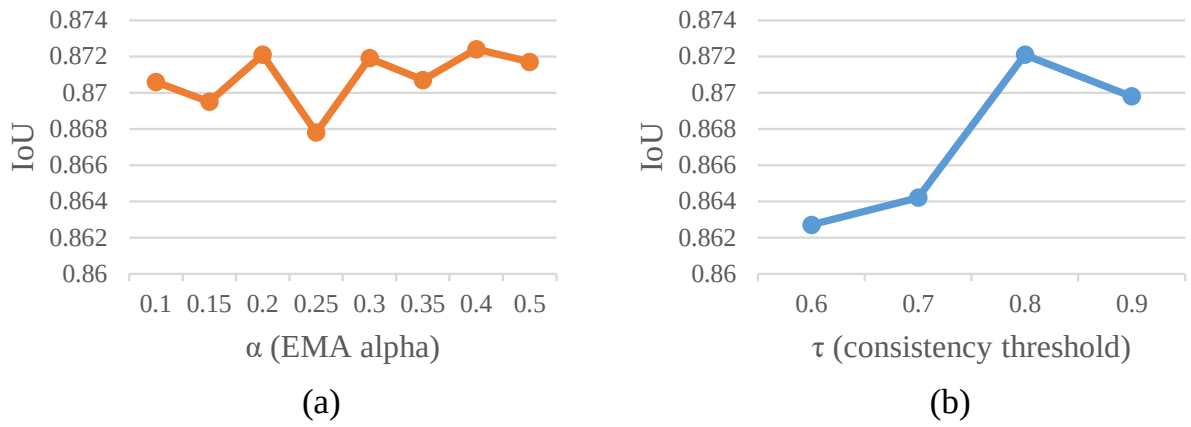


Figure 9: Plots of ablation studies on Scribble2Label. (a) Various EMA Alpha α for a prediction ensemble process. (b) Varying the consistency threshold to measure whether the generated label of the unscribble area is reliable through prediction ensembling. The metric used is Intersection over Union (IoU).

Ablation studies on hyper-parameters

We include an extensive ablation studies on hyper-parameters. The dataset in this experiments was DSB-Fluo and manually generated scribbles are used for label, which is generated in the same rule as 4.1. Default hyper-paramerts used aree as follows: Consistnecy Threshold $\tau = 0.8$; EMA Alpha $\alpha = 0.2$; Ensemble Momentum $\gamma = 5$; $\mathcal{L}_{up}$'s weight $\lambda = 0.5$; and warm-up epoch $E_W = 100$. Consistency Threshold τ and EMA alpha α depend on the experimental setting. The dataset is split into training, validation with ratios of 80%, 20%, respectively.

Fig. 9 shows the quantitative comparison of our method with various EMA alpha α and confidence threshold τ settings. When conducting the experiment with various EMA alpha α, we were able to confirm that the segmentation result was stable. A large EMA alpha α means that adds more weight to relatively recent results, and yet previous results can supplement this, giving robust results to various α settings. This means that sophisticated work is not necessary to set the proper EMA alpha α value. It reduces the huge computation costs. In various the consistency threshold settings τ, we can observe stable performance from a certain reasonable value. This means that unreliable label filtering through prediction ensembling works well.

V Conclusion

In this thesis, `Scribble2Label`, a simple but effective scribble-supervised learning method that combines pseudo-labeling and label-filtering with consistency, is proposed. Unlike the existing methods, `Scribble2Label` demonstrates highly-accurate segmentation performance on various datasets and at different levels of scribble detail without extra segmentation processes or additional model parameters. The proposed method can effectively avoid time-consuming and labor-intensive manual label generation, which is a major bottleneck in image segmentation.

An interesting result found in this work is that a few scribble labels can produce meaningful cell segmentation results. In cell segmentation, cells have many similarities in shape and texture, so the neural network can effectively learn how to recognize. This further demonstrates the possibility of segmentation with a smaller amount of label, or without labels, which is self-supervision. In `Scribble2Label`, there are several hyper-parameters, but the experimental results show that the performance is not sensitive to these values, indicating that there is no need to be particularly careful adjustment. This can significantly reduce the cost of the computation and extend the application range.

In the future, the proposed method will be extended in more general problem settings other than cell segmentation, including semantic and instance segmentation in images and videos. Developing automatic label generation for the segmentation of more complicated biological features, such as tumor regions in histopathology images and mitochondria in nano-scale cell images, is another interesting future research direction. What's interesting is that most of the inconsistent pixels are located on the edge of the cell. Using this information, modeling specialized in instance segmentation is also one of the interests.

References

[1] C. Demir and B. Yener, "Automated cancer diagnosis based on histopathological images: a systematic survey," *Rensselaer Polytechnic Institute, Tech. Rep*, 2005.

[2] T. M. Quan, D. G. Hildebrand, and W.-K. Jeong, "Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics," *arXiv preprint arXiv:1612.05360*, 2016.

[3] P. D. Caie, R. E. Walls, A. Ingleston-Orme, S. Daya, T. Houslay, R. Eagle, M. E. Roberts, and N. O. Carragher, "High-content phenotypic profiling of drug response signatures across distinct cancer cells," *Molecular cancer therapeutics*, vol. 9, no. 6, pp. 1913–1926, 2010.

[4] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[5] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin *et al.*, "Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.

[6] X. Yang, H. Li, and X. Zhou, "Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 11, pp. 2405–2414, 2006.

[7] M. Maška, O. Daněk, S. Garasa, A. Rouzaut, A. Muñoz-Barrutia, and C. Ortiz-de Solorzano, "Segmentation and shape tracking of whole fluorescent cells based on the Chan–Vese model," *IEEE transactions on medical imaging*, vol. 32, no. 6, pp. 995–1006, 2013.

[8] C. Li, J. Liu, and M. D. Fox, "Segmentation of edge preserving gradient vector flow: An approach toward automatically initializing and splitting of snakes," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 162–167.

[9] F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," *IEEE transactions on medical imaging*, vol. 35, no. 2, pp. 550–566, 2015.

[10] A. Arbelle and T. R. Raviv, "Microscopy cell segmentation via convolutional LSTM networks," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1008–1012.

[11] J. Yi, P. Wu, Q. Huang, H. Qu, B. Liu, D. J. Hoeppner, and D. N. Metaxas, "Multi-scale cell instance segmentation with keypoint graph based bounding boxes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 369–377.

[12] K. Nishimura, R. Bise *et al.*, "Weakly Supervised Cell Instance Segmentation by Propagating from Detection Response," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 649–657.

[13] H. Qu, P. Wu, Q. Huang, J. Yi, G. M. Riedlinger, S. De, and D. N. Metaxas, "Weakly supervised deep nuclei segmentation using points annotation in histopathology images," in *International Conference on Medical Imaging with Deep Learning*, 2019, pp. 390–400.

[14] I. Yoo, D. Yoo, and K. Paeng, "PseudoEdgeNet: Nuclei Segmentation only with Point Annotations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 731–739.

[15] D. Lin, J. Dai, J. Jia, K. He, and J. Sun, "Scribblesup: Scribble-supervised convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3159–3167.

[16] Y. B. Can, K. Chaitanya, B. Mustafa, L. M. Koch, E. Konukoglu, and C. F. Baumgartner, "Learning to segment medical images with scribble-supervision alone," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 236–244.

[17] M. Tang, F. Perazzi, A. Djelouah, I. Ben Ayed, C. Schroers, and Y. Boykov, "On regularized losses for weakly-supervised cnn segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 507–522.

[18] B. Wang, G. Qi, S. Tang, T. Zhang, Y. Wei, L. Li, and Y. Zhang, "Boundary perception guidance: a scribble-supervised semantic segmentation approach," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. AAAI Press, 2019, pp. 3663–3669.

[19] W. Wu, H. Qi, Z. Rong, L. Liu, and H. Su, "Scribble-Supervised Segmentation of Aerial Building Footprints Using Adversarial Learning," *IEEE Access*, vol. 6, pp. 58 898–58 911, 2018.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4981–4990.

[22] X. Tian, K. Xu, X. Yang, B. Yin, and R. W. Lau, "Weakly-supervised salient instance detection," *arXiv preprint arXiv:2009.13898*, 2020.

[23] G. French, T. Aila, S. Laine, M. Mackiewicz, and G. Finlayson, "Semi-supervised semantic segmentation needs strong, high-dimensional perturbations," *arXiv preprint arXiv:1906.01916*, 2019.

[24] M. Kim and H. Byun, "Learning texture invariant representation for domain adaptation of semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 975–12 984.

[25] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.

[26] Wikipedia contributors, "Watershed (image processing) — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=Watershed_(image_processing)&oldid=960042704, 2020, [Online; accessed 23-November-2020].

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] T. Lavergne and F. Yvon, "Learning the structure of variable-order crfs: a finite-state perspective," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 433–439.

[30] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2019, pp. 5050–5060.

[31] K. Sohn, D. Berthelot, C.-L. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *arXiv preprint arXiv:2001.07685*, 2020.

[32] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, 2013, p. 2.

[33] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[34] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.

[35] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: Learning to filter noisy labels with self-ensembling," *arXiv preprint arXiv:1910.01842*, 2019.

[36] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," *arXiv preprint arXiv:1908.03265*, 2019.